Is the Jeffreys' scale a reliable tool for Bayesian model comparison in cosmology?

Savvas Nesseris and Juan García-Bellido*
Instituto de Física Teórica UAM-CSIC, Universidad Autonóma de Madrid, Cantoblanco, 28049 Madrid, Spain
(Dated: October 30, 2012)

We are entering an era where progress in cosmology is driven by data, and alternative models will have to be compared and ruled out according to some consistent criterium. The most conservative and widely used approach is Bayesian model comparison. In this paper we explicitly calculate the Bayes factors for all models that are linear with respect to their parameters. We do this in order to test the so called Jeffreys' scale and determine analytically how accurate its predictions are in a simple case where we fully understand and can calculate everything analytically. We also discuss the case of nested models, e.g. one with M_1 and another with $M_2 \supset M_1$ parameters and we derive analytic expressions for both the Bayes factor and the Figure of Merit, defined as the inverse area of the model parameter's confidence contours. With all this machinery and the use of an explicit example we demonstrate that the threshold nature of Jeffreys' scale is not a "one size fits all" reliable tool for model comparison and that it may lead to biased conclusions. Furthermore, we discuss the importance of choosing the right basis in the context of models that are linear with respect to their parameters and how that basis affects the parameter estimation and the derived constraints.

I. INTRODUCTION

Model comparison is at the forefront of modern science, especially in an age of huge datasets and several competing theories. Cosmology has entered an era where large amounts of data will be flowing in from CMB and LSS experiments like Planck [1], BOSS [2], DES [3], COrE [4], Euclid [5], etc. This clearly raises a fundamental question: Given some cosmological observations in the form of data and some cosmological models that may depend on one or more variables, how does one choose the best model? The reason for asking this is quite obvious. Perhaps the models correspond to predictions of different and competing fundamental theories that may explain a range of phenomena. One such example is the plethora of different Dark Energy and Modified Gravity models (see e.g. Ref. [6] for details) that fit the current cosmological observations more or less equally well with General Relativity, at least within the range of a few sigmas.

A common way to answer this question has been by using Bayesian statistics, see Refs. [7–10]. For instance, the usual method of comparing minimum χ^2 per effective degree of freedom normally misses the point and is not very decisive. Other methods to decide which model gives the best description, given the data, include various Information Criteria, e.g. Akaike [11] and Bayesian [12], which are more or less justified, see [13], and normally do not compare well among each other.

On the other hand, the Bayesian evidence is based on Bayes theorem, see Refs. [14],[15] for in-depth reviews, which relates the posterior distribution $\mathcal{P}(u, \mathcal{M}|\mathbf{D})$ for the parameters u of the model $\mathcal{M}$ given the data $\mathbf{D}$, in terms of the likelihood distribution function $\mathcal{L}(\mathbf{D}|u, \mathcal{M})$ within a given set of priors $\pi(u, \mathcal{M})$

$$\mathcal{P}(u, \mathcal{M}|\mathbf{D}) = \frac{\mathcal{L}(\mathbf{D}|u, \mathcal{M})\, \pi(u, \mathcal{M})}{E(\mathbf{D}|\mathcal{M})}\,, \tag{1.1}$$

where the likelihood can be obtained from $\mathcal{L}(\mathbf{D}|u, \mathcal{M}) = \exp(-\chi^2(u)/2)$. Here E is the Bayesian evidence, i.e. the average likelihood over the priors,

$$E(\mathbf{D}|\mathcal{M}) = \int du\ \mathcal{L}(\mathbf{D}|u, \mathcal{M})\, \pi(u, \mathcal{M})\,, \tag{1.2}$$

or roughly, the probability of the data being true given the model, integrated over the whole parameter range u as defined by the priors. The comparison of the models proceeds as the ratio of this quantity evaluated for the different models

$$B_{ij} \equiv \frac{E(\mathbf{D}|\mathcal{M}_i)}{E(\mathbf{D}|\mathcal{M}_j)}\,. \tag{1.3}$$

*Electronic address: savvas.nesseris@uam.es, juan.garciabellido@uam.es

This expression may naively be considered to provide a mathematical representation of Occam's razor, because more complex models tend to be less predictive, lowering their average likelihood (within the priors) in comparison with simpler, more predictive models. Complex models can only be favored if they are able to provide a significantly improved fit to the data. The Bayes factor (1.3) is then used to give evidence of (i.e. favor) the model $\mathcal{M}_i$ against the model $\mathcal{M}_j$ using the so-called Jeffreys' scale, a particular interpretation of the Bayes factor which strengthens its verdict roughly each time the logarithm $\ln B_{ij}$ increases by one unit, from 0 (undecisive) to greater than 5 (strongly ruled out). More details on the Jeffreys' scale can be found in later sections and specific threshold values in Table II.

In Section II, we explicitly calculate the Bayes factors for all models that are linear with respect to their parameters. In Section III we discuss the case of nested models, e.g. one with M_1 and another with M_2 parameters and we derive analytic expressions for both the Bayes factor while in Section IV we discuss the same problem for the Figure of Merit. With all this machinery and the use of the explicit example we demonstrate in Section III that the Jeffreys' scale is not a "one size fits all" reliable tool for model comparison, contrary to the common belief by many people.

II. GENERAL LINEAR LEAST-SQUARES FITTING

A. Minimization

In this section we will briefly discuss the case of the general linear least squares fitting. Given some data that consist of N measurements (x_i, y_i, σ_i), where $i = (1, 2, ..., N)$, and a model which is a linear combination of M functions,

$$y(x) = \sum_{i=0}^{M-1} a_i X_i(x)\,, \tag{2.1}$$

then the fitness of the model with respect to the data and the parameters a_i, for $i = (0, 1, ..., M-1)$, can be found by calculating the $\chi^2(\vec{a})$ defined as [16], [17]

$$\begin{aligned}\chi^2(\vec{a}) &\equiv \sum_{i=1}^{N}\left(\frac{y_i - y(x_i;a_j)}{\sigma_i}\right)^2 \\ &= \sum_{i=1}^{N}\left(\frac{y_i - \sum_{j=0}^{M-1} a_j X_j(x_i)}{\sigma_i}\right)^2\end{aligned} \tag{2.2}$$

The base $X_i(x)$ can be any set of M functions, e.g. monomials like $\{x^i\}_{i=0}^{M-1}$ or Chebyshev polynomials $\{T_i(x)\}_{i=0}^{M-1}$, of order M. The latter are a set of orthogonal polynomials that can act as a base of functions with the property that when $x \in [-1, 1]$ they have the smallest maximum deviation from the true function at any given order M. The first few Chebyshev polynomials are $T_0(x) = 1,\ T_1(x) = x,\ T_2(x) = -1 + 2x^2,\ T_3(x) = -3x + 4x^3$. When $x \in [-1, 1]$, the variable x can be written as $x = \cos(\theta)$ and the polynomials can also be expressed as $T_n(\cos(\theta)) = \cos(n\theta) = \cos(n \arccos(x))$, which implies that $|T_n(x)| \leq 1$. Since in general our data will not be in the range $[-1, 1]$, we can normalize x by using $\tilde{x} = \frac{2x}{x_{max}} - 1$ and using instead the basis $T_n(\tilde{x}) \equiv T_n(\frac{2x}{x_{max}} - 1)$, where x_{max} is the maximum value of the N data x_i. From now on we will assume that x has been normalized and we will drop the tilde on x. Finally, we will mostly follow the notation of Ref. [16].

The best-fit parameters of the model can be found by minimizing the $\chi^2(\vec{a})$ of Eq. (2.2) with respect to the parameters a_j. This is done by demanding that the derivatives of $\chi^2(\vec{a})$ are equal to zero at the minimum, i.e. $\partial_j \chi^2(a_k) = 0$. Then, this gives [16], [17]

$$\begin{aligned}\partial_j \chi^2(a_k) &= \sum_{i=1}^{N}\left(\sum_{k=0}^{M-1}(-2)\frac{\partial a_k}{\partial a_j}\frac{X_k(x_i)}{\sigma_i}\right)\left(\frac{y_i - \sum_{m=0}^{M-1} a_m X_m(x_i))}{\sigma_i}\right) \\ &= (-2)\sum_{i=1}^{N}\frac{X_j(x_i)}{\sigma_i}\left(\frac{y_i - \sum_{m=0}^{M-1} a_m X_m(x_i)}{\sigma_i}\right) = 0 \\ &\text{or} \\ \left(\sum_{i=1}^{N}\frac{y_i X_j(x_i)}{\sigma_i^2}\right) &= \left(\sum_{k=0}^{M-1}\sum_{i=1}^{N}\frac{X_j(x_i)X_k(x_i)}{\sigma_i^2}a_k\right).\end{aligned} \tag{2.3}$$

If we define the Fisher Matrix F_{ij} and the constant vector β_j as

$$F_{ij} = \frac{1}{2}\partial_{ij}\chi^2|_{\rm min} = \sum_{k=1}^{N}\frac{X_i(x_k)X_j(x_k)}{\sigma_k^2} = \text{ const.} \tag{2.4}$$

$$\beta_j = \sum_{i=1}^{N}\frac{y_i X_j(x_i)}{\sigma_i^2} = \text{ const.}, \tag{2.5}$$

where as usual $j = (0, 1, ..., M-1)$, then Equation (2.3) can be rewritten in matrix form and easily solved for the best-fit parameters $\vec{a}_{\rm min}$ as

$$\begin{aligned} \beta_j &= F_{jk}\ a_{k,\rm min} \\ a_{k,\rm min} &= F_{kj}^{-1}\beta_j = C_{kj}\beta_j, \end{aligned} \tag{2.6}$$

where $C_{kj} \equiv F_{kj}^{-1}$ is the covariance matrix. If we define the parameter $S_y \equiv \sum_{i=1}^{N} y_i^2/\sigma_i^2$, then the χ^2 at the minimum can be written as

$$\begin{aligned} \chi^2_{\rm min} &= S_y - C_{ij}\beta_i\beta_j \\ &= S_y - F_{ij}a_{i,\rm min}a_{j,\rm min}\,. \end{aligned} \tag{2.7}$$

Finally, the 1σ errors on the best-fit parameters are given by the diagonal elements of the covariance matrix

$$\sigma(a_k)^2 = C_{kk} \tag{2.8}$$

At this point we should note that by doing a Taylor expansion around the minimum the $\chi^2(\vec{a})$ can also be written as

$$\chi^2(\vec{a}) = \chi^2_{\rm min} + (a - a_{\rm min})_i\ F_{ij}\ (a - a_{\rm min})_j, \tag{2.9}$$

since the first derivatives at the minimum are by definition equal to zero and all higher derivatives $\partial^n_{i_1\ldots i_n}\chi^2(\vec{a})$ for $n \geq 3$ vanish like in our model.

B. Change of basis

At this point we should note that the functional form of the results of Eqs. (2.6)-(2.9) is completely independent of the basis used, regardless of it being some combination of polynomials (monomials or Chebyshev polynomials) or something more complicated, e.g. $\sin(n\ x)$ etc. For example, in the case of the monomials the Fisher matrix is equal to

$$F_{ij} = \frac{1}{2}\partial_{ij}\chi^2|_{\rm min} = \sum_{k=1}^{N}\frac{x_k^i x_k^j}{\sigma_k^2}, \tag{2.10}$$

for $i = (0, 1, ..., M-1)$. If at this point we define a constant

$$S_n \equiv \sum_{k=1}^{N}\frac{x_k^n}{\sigma_k^2}, \tag{2.11}$$

where for example $S_0 = \sum_{k=1}^{N}\frac{1}{\sigma_k^2}$, $S_1 = \sum_{k=1}^{N}\frac{x_k}{\sigma_k^2}$, $S_2 = \sum_{k=1}^{N}\frac{x_k^2}{\sigma_k^2}$ and so on, then in the case that $M = 3$ the Fisher matrix will be given by

$$F_{ij} = \begin{pmatrix} S_0 & S_1 & S_2 \\ S_1 & S_2 & S_3 \\ S_2 & S_3 & S_4 \end{pmatrix}, \tag{2.12}$$

where the constants S_n will only depend on the data.

For the Chebyshev polynomials the Fisher matrix is equal to

$$F_{ij} = \frac{1}{2}\partial_{ij}\chi^2|_{\rm min} = \sum_{k=1}^{N}\frac{T_i(x_k)T_j(x_k)}{\sigma_k^2}, \tag{2.13}$$

For example, in the case that $M = 3$ the Fisher matrix will be given by

$$F_{ij} = \begin{pmatrix} \sum_{k=0}^{N} \frac{T_0(x_k)^2}{\sigma_k^2} & \sum_{k=0}^{N} \frac{T_0(x_k)T_1(x_k)}{\sigma_k^2} & \sum_{k=0}^{N} \frac{T_0(x_k)T_2(x_k)}{\sigma_k^2} \\ \sum_{k=0}^{N} \frac{T_0(x_k)T_1(x_k)}{\sigma_k^2} & \sum_{k=0}^{N} \frac{T_1(x_k)^2}{\sigma_k^2} & \sum_{k=0}^{N} \frac{T_1(x_k)T_2(x_k)}{\sigma_k^2} \\ \sum_{k=0}^{N} \frac{T_0(x_k)T_2(x_k)}{\sigma_k^2} & \sum_{k=0}^{N} \frac{T_1(x_k)T_2(x_k)}{\sigma_k^2} & \sum_{k=0}^{N} \frac{T_2(x_k)^2}{\sigma_k^2} \end{pmatrix}, \tag{2.14}$$

which obviously is constant and will only depend on the data at hand[1]. Similar expressions can be derived for other cases as well.

In general, if we change basis from $X_i(x)$ to some new *polynomial* basis $\tilde{X}_i(x)$ of the same *order*, then we will have to replace the parameters a_k with some new parameters $\tilde{a}_k$, assuming that we still have the same number of linearly dependent parameters M. However, the function $y(x)$ will still be the same, so

$$\begin{aligned} y(x) &= \sum_{i=0}^{M-1} a_i X_i(x) \\ &= \sum_{i=0}^{M-1} \tilde{a}_i \tilde{X}_i(x)\,, \end{aligned} \tag{2.15}$$

Let one of the two bases satisfy an orthogonality relation with a weight $w(x)$:

$$\int_{x_1}^{x_2} dx\ w(x)\ \tilde{X}_i(x)\tilde{X}_j(x) = c_i \delta_{ij} \tag{2.16}$$

for some constants c_i. Then, by using Eq. (2.15) and the orthogonality relation we can derive a transformation Λ between the two sets of parameters and for the other quantities of interest:

$$a_i = \Lambda_{ij}^{-1} \tilde{a}_j \tag{2.17}$$

$$\beta_i = \Lambda_{ij}^{T} \tilde{\beta}_j \tag{2.18}$$

$$F_{ij} = \Lambda_{ik}^{T} \tilde{F}_{kl} \Lambda_{lj} \tag{2.19}$$

$$\tilde{C}_{ij} = \Lambda_{ik} C_{kl} \Lambda_{lj}^{T} \tag{2.20}$$

Finally, as it can easily be seen from Eq. (2.7) the χ^2_{min} as expected is invariant under the transformation Λ.

For example, we will now consider the change of basis from the monomials to the Chebyshev polynomials. In this case we will have

$$\sum_{n=0}^{M-1} a_n x^n = \sum_{n=0}^{M-1} \tilde{a}_n T_n \left(\frac{2x}{x_{max}} - 1 \right), \tag{2.21}$$

Remembering the fact that the Chebyshev polynomials satisfy the orthogonality relation

$$\int_{-1}^{1} dz \frac{T_k(z) T_n(z)}{\sqrt{1-z^2}} = k_n \delta_{nk} \tag{2.22}$$

where $k_0 = \pi$ and $k_n = \pi/2$ for $n \geq 1$. Then we can multiply Eq. (2.21) with the appropriate factors $\frac{T_n(z)}{\sqrt{1-z^2}}$, where $z = \frac{2x}{x_{max}} - 1$ and by integrating both sides over $z \in [-1, 1]$ we get the transformation between the two sets of parameters as

$$\tilde{a}_n = \Lambda_{kn} a_k \tag{2.23}$$

---

[1] At this point we should remind the reader of our shorthand convention that x is normalized, so for example by writing $T_0(x_k)$ we actually imply $T_0\left(\frac{2x_k}{x_{max}} - 1\right)$.

TABLE I: The determinant of the transformation matrix Λ_{kn} for various combinations of polynomials including the monomials x^n, the Legendre polynomials $P_n(x)$ and the Chebyshev polynomials $T_n(x)$. These values are particularly important in the estimation of the Figure of Merit, as shown in a later section.

$\lvert\Lambda\rvert$	Monomials	Legendre	Chebyshev
Monomials	1	$\frac{x_m^3}{12}$	$\frac{x_m^3}{16}$
Legendre	$\frac{12}{x_m^3}$	1	$\frac{3}{4}$
Chebyshev	$\frac{16}{x_m^3}$	$\frac{4}{3}$	1

where the constant matrix Λ is given by

$$\Lambda_{kn} = k_n^{-1} \int_{-1}^{1} dz \frac{\left(\frac{x_{max}}{2}(z+1)\right)^k T_n(z)}{\sqrt{1-z^2}} \tag{2.24}$$

In Table I we show the determinant of the transformation matrix Λ_{kn} for various combinations of polynomials including the monomials x^n, the Legendre polynomials $P_n(x)$ and the Chebyshev polynomials $T_n(x)$. These values are particularly important in the estimation of the Figure of Merit, as shown in a later section. Finally, we should stress that these results are only valid for a transformation from the original *polynomial* basis $X(x)$ with a set of M linearly dependent parameters a_i to a new set with the same number of linearly dependent parameters M and a *polynomial* basis $\tilde{X}(x)$.

C. Likelihood calculations

After having determined the best-fit parameters in the previous section, we will now define the likelihood for our model. This is defined as [16]:

$$\mathcal{L} = \mathcal{N} \exp\left(-\chi^2(\vec{a})/2\right), \tag{2.25}$$

where the parameter $\mathcal{N}$ can be found by normalizing the likelihood, ie integrating it over all parameters. In our case this means :

$$\int \mathcal{L} d\vec{a} = \int_{-\infty}^{\infty} \mathcal{N} \exp\left(-\chi^2(\vec{a})/2\right) d\vec{a} \; =$$
$$\mathcal{N} e^{-\chi^2_{\rm min}/2} \int_{-\infty}^{\infty} e^{-1/2(a-a_{\rm min})_i \; F_{ij} \; (a-a_{\rm min})_j} da_0 da_1 ... da_{M-1} \; = \; 1 \tag{2.26}$$

To proceed we now have to rotate the parameters to a basis where they are not correlated with each other. To do so we define a new variable $s_i \equiv D_{ij}\left(a_j - a_{j,\rm min}\right)$, where D_{ij} can be found by decomposing the inverse covariance matrix $F = C^{-1} = D^T D$ by using Cholesky decomposition[2]. Then, we have that $ds_1 ... ds_N = |D|\, df_1 ... df_N$ and the integration can proceed as usual and the normalization can be found. Going to the new basis we have

$$s_i \; \equiv \; D_{ij}\left(a_j - a_{j,\rm min}\right) \tag{2.27}$$
$$ds_1 ... ds_N \; = \; |D|\, da_0 da_1 ... da_{M-1} \tag{2.28}$$
$$|D| \; = \; |F|^{1/2} = |C|^{-1/2} \tag{2.29}$$

where $a_{j,\rm min}$ is to be understood as the value of the jth parameter a_j at its best-fit value (the "minimum"). Then,

[2] Cholesky decomposition can easily be implemented in computer programs such as Mathematica. For example, in the latter the Cholesky decomposition of a matrix $M = D^T D$ is given by $D = CholeskyDecomposition[M]$. This works both symbolically and numerically.

Eq. (2.26) becomes

$$\begin{aligned}\mathcal{N}\, e^{-\chi^2_{\rm min}/2} \int_{-\infty}^{+\infty} e^{-\sum_{i=0}^{M-1} s_i^2/2}\, |D|^{-1} \prod_{i=0}^{M-1} ds_i \;&=\\ \mathcal{N}\, e^{-\chi^2_{\rm min}/2}\, |F|^{-1/2} \prod_{i=0}^{M-1} \int_{-\infty}^{+\infty} e^{-s_i^2/2} ds_i \;&=\\ \mathcal{N}\, e^{-\chi^2_{\rm min}/2}\, |F|^{-1/2}\, (2\pi)^{M/2} \;&=\; 1,\end{aligned} \tag{2.30}$$

and finally,

$$\mathcal{N} = e^{\chi^2_{\rm min}/2}\, |F|^{1/2}\, (2\pi)^{-M/2}, \tag{2.31}$$

Unsurprisingly, the resulting normalized likelihood now becomes

$$\begin{aligned}\mathcal{L} \;&=\; \frac{|F|^{1/2}}{(2\pi)^{M/2}} \exp\left(-(\chi^2(\vec{a}) - \chi^2_{\rm min})/2\right), \\ &=\; \frac{1}{(2\pi)^{M/2}\, |C|^{1/2}} \exp\left(-(\chi^2(\vec{a}) - \chi^2_{\rm min})/2\right),\end{aligned} \tag{2.32}$$

where in the last line we used the fact that $|F| = |C|^{-1}$.

III. BAYESIAN MODEL COMPARISON

A. The Jeffreys' scale revisited

In this section we will present results related to the Bayes factor B_{ij}, where (see [18] and references there-in), in the context of our simple model. In this case, the Bayes factor B_{ij} can be written as

$$B_{ij} \equiv \frac{L(M_i)}{L(M_j)} \tag{3.1}$$

where $L(M_i)$ denotes the probability $p(D|M_i)$, called likelihood for the model M_i, to obtain the data D if the model M_i is the true one. Generally, $L(M_i)$ is defined as:

$$L(M_i) \equiv p(D|M_i) = \int da \cdot p(a|M_i)\mathcal{L}_i(a) \tag{3.2}$$

for models with one free parameter and where $p(a|M_i)$ is the prior probability for the parameter a. Also, $\mathcal{L}_i(a)$ is the likelihood for the parameter a in the model and

$$\mathcal{L}_i(a) \equiv e^{-\chi^2(a)/2} \tag{3.3}$$

In the case that a has flat prior probabilities, that is we have no prior information on a besides that it lies in some range $[a, a+\Delta a]$ then $p(a|M_i) = \frac{1}{\Delta a}$ and

$$L(M_i) = \frac{1}{\Delta a} \int_a^{a+\Delta a} da e^{-\chi^2(a)/2} \tag{3.4}$$

The reason why we did not use the normalized likelihood is that the evidence $L(M_i)$ has to be *dimensionless* in general. As it can be seen by inspecting Eq. (2.25), this is achieved since the units of the prior Δa will cancel out with the ones from the infinitesimal quantity da in the integral. However, this would not happen had we included the normalization constant $\mathcal{N}$.

Of course, all this can be easily generalized for models having more than one parameter as follows

$$L(M_i) = \left(\prod_{j=0}^{M-1} \frac{1}{\Delta a_j}\right) \int_{\vec{a}}^{\vec{a}+\Delta\vec{a}} e^{-\chi^2(\vec{a})/2} \cdot d\vec{a} \tag{3.5}$$

TABLE II: The values of both the linear and the logarithmic Jeffreys' scale, and the AIC and BIC criteria. For references on these values check the text.

B_{ij}	$\ln B_{ij}$	Δ(AIC)	Δ(BIC)	Evidence
< 3	< 1.1	< 2	< 2	Weak
< 20	< 3	< 6	< 6	Definite
< 150	< 5	< 10	< 10	Strong
> 150	> 5	> 10	> 10	Very Strong

where M is the total number of parameters and the integration over $d\vec{a} \equiv \prod_{j=0}^{M-1} da_j = da_0 da_1 ... da_{M-1}$ is assumed to be multidimensional in general. Also, we will consider Gaussian priors of the form:

$$Pr(\vec{a}) = \frac{|H_{ij}|^{1/2}}{(2\pi)^{M/2}}\, e^{-(a-a_{\rm pr})_i\ H_{ij}\ (a-a_{\rm pr})_j/2}. \tag{3.6}$$

where the M priors are centered around the values $a_{\rm pr,i}$ and H_{ij} is their inverse covariance. Also, we have properly normalized the gaussian priors to unity, such that $\int_{-\infty}^{+\infty} Pr(\vec{a}) d\vec{a} = 1$.

The interpretation of the Bayes factor B_{ij} is that [18] when $1 < B_{ij} < 3$ there is evidence against M_j when compared with M_i, but it is only worth a bare mention. When $3 < B_{ij} < 20$ the evidence against M_j is definite but not strong. For $20 < B_{ij} < 150$ the evidence is strong and for $B_{ij} > 150$ it is very strong. For handy reference we include the values of both the linear and the logarithmic Jeffreys' scale in Table II. Jeffreys in his seminal paper [7, 19] provides somewhat different but in general consistent values. Several examples of the use of the Jeffreys' scale in cosmology and astronomy can by found in [18, 20–22] and references there-in.

B. The Bayesian evidence

1. Gaussian priors

Using the machinery of the previous sections we will now calculate the Bayesian evidence of Eq. (3.5) in the case of the Gaussian priors. Using Eqs. (3.2) and (3.6) we have:

$$\begin{aligned} B_1 &= \int_{-\infty}^{+\infty} e^{-\chi^2(\vec{a})/2} Pr(\vec{a}) \cdot d\vec{a} \\ &= \frac{|H_{ij}|^{1/2}}{(2\pi)^{M/2}} e^{-\chi^2_{\rm min}/2} \int_{-\infty}^{+\infty} e^{-(a-a_{\rm min})_i\ F_{ij}\ (a-a_{\rm min})_j/2 - (a-a_{\rm pr})_i\ H_{ij}\ (a-a_{\rm pr})_j/2} \cdot d\vec{a} \end{aligned} \tag{3.7}$$

At this point we can introduce a new matrix G_{ij} and constants c and $a_{1,i}$ such that:

$$G_{ij} = F_{ij} + H_{ij} \tag{3.8}$$

$$a_{1,i} = (a_{k,\rm min} F_{kj} + a_{k,\rm pr} H_{kj})\, G^{-1}_{ij} \tag{3.9}$$

$$c = (a_{\rm min} - a_1)_i\ F_{ij}\ (a_{\rm min} - a_1)_j + (a_{\rm pr} - a_1)_i\ H_{ij}\ (a_{\rm pr} - a_1)_j \tag{3.10}$$

$$(a-a_1)_i\ G_{ij}\ (a-a_1)_j + c = (a-a_{\rm min})_i\ F_{ij}\ (a-a_{\rm min})_j + (a-a_{\rm pr})_i\ H_{ij}\ (a-a_{\rm pr})_j \tag{3.11}$$

Obviously, when the best-fit and the priors are centered in the same point, ie $a_{i,\rm min} = a_{i,\rm pr}$, then we have that $a_{1,i} = a_{i,\rm min} = a_{i,\rm pr}$ and $c = 0$. By using Eqs. (3.11) we can proceed with the calculation of (3.7) as usual:

$$\begin{aligned} B_1 &= \frac{|H_{ij}|^{1/2}}{(2\pi)^{M/2}} e^{-\chi^2_{\rm min}/2} \int_{-\infty}^{+\infty} e^{-(a-a_1)_i\ G_{ij}\ (a-a_1)_j/2 - c/2} \cdot d\vec{a} \\ &= \frac{|H_{ij}|^{1/2}}{(2\pi)^{M/2}} e^{-\chi^2_{\rm min}/2 - c/2} \frac{(2\pi)^{M/2}}{|G_{ij}|^{1/2}} \\ &= e^{-\chi^2_{\rm min}/2 - c/2} \frac{|H_{ij}|^{1/2}}{|G_{ij}|^{1/2}} \\ &= e^{-\chi^2_{\rm min}/2 - c/2} |I_M + H^{-1} F|^{-1/2} \end{aligned} \tag{3.12}$$

where in the second line we used Eqs. (2.27)-(2.29), in the last line we used the well known matrix identity $|X+A| = |X||I_M + X^{-1}A|$ where I_M is the $M \times M$ unit matrix, and finally we used the fact that $G_{ij} = H_{ij} + F_{ij}$. By using the matrix identity $|A| = \frac{1}{2}\left(tr(A)^2 - tr(A^2)\right)$, where $tr(A)$ is the trace of the matrix A, we can express the determinant of a sum of the unit matrix I_M and a matrix B as

$$\begin{aligned} |I_M + B| &= \frac{1}{2}\left(\mathrm{tr(I_M + B)^2 - tr((I_M + B)\cdot(I_M + B))}\right) \\ &= \frac{M(M-1)}{2} + (M-1)\mathrm{tr(B)} + \mathrm{|B|} \end{aligned} \tag{3.13}$$

and then, the expression $|I_M + H^{-1}F|$ can be expanded to

$$|I_M + H^{-1}F| = \frac{M(M-1)}{2} + (M-1)\mathrm{tr(H^{-1}F)} + \mathrm{|H^{-1}||F|} \tag{3.14}$$

Then the Bayes factor can be written as

$$\begin{aligned} B_{12} &= e^{-\Delta\chi^2_{1,2\mathrm{min}}/2 - \Delta c_{1,2}/2} \frac{|I_{M_1} + H^{-1}_{(1)}F_{(1)}|^{-1/2}}{|I_{M_2} + H^{-1}_{(2)}F_{(2)}|^{-1/2}} \\ &= e^{-\Delta\chi^2_{1,2\mathrm{min}}/2 - \Delta c_{1,2}/2} \left(\frac{\frac{M_1(M_1-1)}{2} + (M_1-1)\mathrm{tr(H^{-1}_{(1)}F_{(1)})} + \mathrm{|H^{-1}_{(1)}||F_{(1)}|}}{\frac{M_2(M_2-1)}{2} + (M_2-1)\mathrm{tr(H^{-1}_{(2)}F_{(2)})} + \mathrm{|H^{-1}_{(2)}||F_{(2)}|}} \right)^{-1/2} \end{aligned} \tag{3.15}$$

where $\Delta\chi^2_{1,2\mathrm{min}} = \chi^2_{1\mathrm{min}} - \chi^2_{2\mathrm{min}}$ and $\Delta c_{1,2} = c_1 - c_2$ are the values of the constant c for the two models. Also, in the last line we have labeled all the different quantities with (1) and (2) to indicate the two models 1 and 2 respectively. Finally, it should be noted that Eq. (3.15) is an *exact* result.

2. Flat priors

Alternatively, we can choose our priors to be top-hat and centered around the best-fit, ie we integrate in the range $[\vec{a}_{\mathrm{min}} - \frac{\Delta\vec{a}}{2}, \vec{a}_{\mathrm{min}} + \frac{\Delta\vec{a}}{2}]$, with our flat top-hat priors being equal to $\Delta\vec{a} = \overrightarrow{const.} \in R^M$, where M is as usual the total number of parameters of the model.

In this case then, by using Eq. (3.5), the evidence B_1 for the model can be written as

$$\begin{aligned} B_1 &= \left(\prod_{j=0}^{M-1} \frac{1}{\Delta a_j}\right) \int_{\vec{a}_{\mathrm{min}} - \Delta\vec{a}/2}^{\vec{a}_{\mathrm{min}} + \Delta\vec{a}/2} e^{-\chi^2(\vec{a})/2} \cdot d\vec{a} \\ &= \left(\prod_{j=0}^{M-1} \frac{1}{\Delta a_j}\right) e^{-\chi^2_{\mathrm{min}}/2} \int_{\vec{a}_{\mathrm{min}} - \Delta\vec{a}/2}^{\vec{a}_{\mathrm{min}} + \Delta\vec{a}/2} e^{-(a-a_{\mathrm{min}})_i \ F_{ij} \ (a-a_{\mathrm{min}})_j/2} \cdot d\vec{a} \end{aligned} \tag{3.16}$$

Using Eqs. (2.9) and (2.27)-(2.29) we can define a transverse of the Fisher matrix, $F^\perp \equiv U^{-1} F U$, where U is the unitary off-diagonal M-dimensional matrix, e.g. for $M = 2$,

$$U = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

as well as the transverse of the Cholesky decomposition of $F^\perp$, denoted with the matrix $D^\perp$. With this, we can write (3.16) as

$$\begin{aligned} B_1 &\approx \left(\prod_{j=0}^{M-1} \frac{1}{\Delta a_j}\right) e^{-\chi^2_{\mathrm{min}}/2} \left|F\right|^{-1/2} (2\pi)^{M/2} \prod_{i=0}^{M-1} \mathrm{erf}\left(\frac{\mathrm{D^\perp_{ii}}\Delta\mathrm{a_i}}{2\sqrt{2}}\right) \\ &= \frac{1}{\mathcal{N}_1 V_{1,prior}} \prod_{i=0}^{M_1-1} \mathrm{erf}\left(\frac{\mathrm{D^{\perp(1)}_{ii}}\Delta\mathrm{a^{(1)}_i}}{2\sqrt{2}}\right), \end{aligned} \tag{3.17}$$

where we have used Eq. (2.31) and we have defined $V_{1,prior} \equiv \prod_{j=0}^{M_1-1} \Delta a_j^{(1)}$ as the "volume" of our priors for this model 1. The function erf(x) is the usual error function defined as $\mathrm{erf(x)} \equiv \frac{2}{\sqrt{\pi}} \int_0^{\mathrm{x}} \mathrm{e}^{-\mathrm{t}^2} \mathrm{dt}$, see Ref. [23] for more details. Then the Bayes factor for two models based on Eq. (2.1), labeled 1 and 2 with a total number of parameters M_1 and M_2, is

$$B_{12} = \frac{\mathcal{N}_2 V_{2,prior}}{\mathcal{N}_1 V_{1,prior}} \cdot \frac{\prod_{i=0}^{M_1-1} \mathrm{erf}\left(\frac{\mathrm{D}_{\mathrm{ii}}^{\perp(1)} \Delta \mathrm{a}_{\mathrm{i}}^{(1)}}{2\sqrt{2}}\right)}{\prod_{i=0}^{M_2-1} \mathrm{erf}\left(\frac{\mathrm{D}_{\mathrm{ii}}^{\perp(2)} \Delta \mathrm{a}_{\mathrm{i}}^{(2)}}{2\sqrt{2}}\right)} . \tag{3.18}$$

Also, in the last line we have labeled all the different quantities with a $(...)^{(1)}$ to indicate model 1. At this point we will now consider two different cases:

- When the priors are much smaller than the errors of the best-fit parameters, ie the arguments of the error functions are small, or $D_{ii}^{\perp} \Delta a_i \ll 1$.
- When the priors are much larger than the errors of the best-fit parameters, ie the arguments of the error functions are large, or $D_{ii}^{\perp} \Delta a_i \gg 1$.

The following expansions of the error function are useful:

$$\begin{aligned} \mathrm{erf(x)} &\approx \frac{2\sqrt{2}\, x}{\sqrt{2\pi}} \left(1 - \frac{x^2}{3}\right) + ... \quad \text{for} \quad x \ll 1 \\ \mathrm{erf(x)} &\approx 1 - \frac{e^{-x^2}}{\sqrt{\pi} x} + ... \quad \text{for} \quad x \gg 1 \end{aligned} \tag{3.19}$$

In the first case (when $x \ll 1$), the evidence B_1 of Eq. (3.17) becomes

$$\begin{aligned} B_1 &\approx e^{-\chi^2_{\mathrm{min}}/2} \prod_j^{M-1} \left(1 - \frac{(D_{jj}^{\perp} \Delta a_j)^2}{4!} + ...\right) \\ &= e^{-\chi^2_{\mathrm{min}}/2} \left(1 - \sum_j^{M-1} \frac{(D_{jj}^{\perp} \Delta a_j)^2}{4!} + ...\right) \end{aligned} \tag{3.20}$$

where we have used the fact that $\prod_{j=0}^{M-1} D_{jj}^{\perp} = |D^{\perp}| = |F|^{1/2}$. Then the Bayes factor can be written as

$$\begin{aligned} B_{12} &\approx e^{-\Delta\chi^2_{1,2\mathrm{min}}/2} \frac{\prod_{j=0}^{M_1-1} \left(1 - \frac{(D_{jj}^{\perp(1)} \Delta a_j^{(1)})^2}{4!} + ...\right)}{\prod_{j=0}^{M_2-1} \left(1 - \frac{(D_{jj}^{\perp(2)} \Delta a_j^{(2)})^2}{4!} + ...\right)} \\ &= e^{-\Delta\chi^2_{1,2\mathrm{min}}/2} \left(1 - \sum_{j=0}^{M_1-1} \frac{(D_{jj}^{\perp(1)} \Delta a_j^{(1)})^2}{4!} + \sum_{j=0}^{M_2-1} \frac{(D_{jj}^{\perp(2)} \Delta a_j^{(2)})^2}{4!} + ...\right) \end{aligned} \tag{3.21}$$

where $\Delta\chi^2_{1,2\mathrm{min}} = \chi^2_{1\mathrm{min}} - \chi^2_{2\mathrm{min}}$. Also, in the last line we have labeled all the different quantities with $(...)^{(1)}$ and $(...)^{(2)}$ to indicate the models 1 and 2 respectively. In this limit the second term of the expression is expected to be close to 1.

In the second case (when $x \gg 1$) the evidence becomes

$$B_1 \approx \left(\prod_{j=0}^{M-1} \frac{1}{\Delta a_j}\right) e^{-\chi^2_{\mathrm{min}}/2} \, |F|^{-1/2} \, (2\pi)^{M/2} \left(1 - \frac{1}{(2\pi)^{1/2}} \sum_{j=0}^{M-1} \frac{e^{-(D_{jj}^{\perp} \Delta a_j)^2/8}}{D_{jj}^{\perp} \Delta a_j / 4}\right) \tag{3.22}$$

where in this limit the last term of the expression is expected to be close to 1. Then the Bayes factor is just $B_{12} = \frac{B_1}{B_2}$.

C. The information criteria

The Akaike information criterion (AIC) is defined as

$$\mathrm{AIC} = -2ln\mathcal{L} + 2M, \tag{3.23}$$

where $\mathcal{L}$ is the maximum likelihood and M the number of parameters of the model [11], [13]. Between two models M_1 and M_2, the better of the two, is the one which minimizes the AIC. Typically, models with too few parameters give a poor fit to the data and hence have a low log-likelihood or χ^2, while models with too many are penalized by the last term.

On the other hand, the Bayesian information criterion (BIC) is defined as

$$\mathrm{BIC} = -2ln\mathcal{L} + 2M \ln N, \tag{3.24}$$

where N is the number of datapoints used in the fit. In either case however, the absolute value of the criterion is irrelevant and only the difference between the two models is important. Usually, differences $\Delta(\mathrm{AIC/BIC}) < 2$ represent weak evidence, differences between 2 and 6 represent positive evidence, $6-10$ strong evidence, and > 10 very strong evidence [24]. At this point the astute reader might have noticed in Table II that the threshold values for the two information criteria are twice a big as in the case of the logarithmic Bayes factor. The reason for this of course is the factors of 2 that appear in Eqs. (3.23) and (3.24).

D. Analysis

As it can be seen from Eqs. (3.18) and (3.21), in both cases (gaussian and flat priors) the Bayes factor for this class of models can be written as

$$B_{12} = e^{-\Delta\chi^2_{1,2\mathrm{min}}/2} \cdot G(M_1, M_2) \tag{3.25}$$

where the function $G(M_1, M_2)$ contains all the extra information of the models via their covariance matrices. Then the logarithmic Bayes factor can be written as

$$\ln(B_{12}) = -\Delta\chi^2_{1,2\mathrm{min}}/2 + \ln(G(M_1, M_2)) \tag{3.26}$$

For example, in this order of the approximation and in the case of the flat priors Eq. (3.21) gives

$$\ln(B_{12}) = -\Delta\chi^2_{1,2\mathrm{min}}/2 - \sum_{j=0}^{M_1-1} \frac{(D^{\perp(1)}_{jj}\Delta a^{(1)}_j)^2}{4!} + \sum_{j=0}^{M_2-1} \frac{(D^{\perp(2)}_{jj}\Delta a^{(2)}_j)^2}{4!} + ... \tag{3.27}$$

As it can be seen, Eq. (3.27) does not only contain the difference between the χ^2_{min} of the two models but also contains information on their covariances via the the last two terms. Clearly, these two terms *depend strongly* on the data and the model at hand, thus introducing a further complexity in the model comparison.

To back up our claims we also present an explicit example. First, we created a set of 31 mock data points (x_i, y_i, σ_{y_i}) with noise in the range $x \in [0.025, 1.55]$ based on the same model as in [25]

$$f(x) = a + (x-b)\exp(-c\ x^2), \tag{3.28}$$

where the parameters (a, b, c) have the values $(0.25, 0.25, 0.5)$ respectively. The choice of the model was completely ad hoc, except for the requirement to be well behaved (smooth) and that it exhibits some interesting features, such as a maximum at some point x. Then, we *fit these data with two polynomials* that have a different number of parameters M_1 and M_2, e.g. $M_2 > M_1$ or $M_1 > M_2$. In other words, our model comparison is done between the models M_1 and M_2 and not between $f(x)$ of Eq. (3.28) and a polynomial.

In Fig. 1 we show the best-fit χ^2 as a function of the number of parameters M (top left), the best-fit χ^2 per degree of freedom $N-M$ as a function of the number of parameters M (top right), the difference in the best-fit χ^2 between two models wtih parameters $M_1 = M+1$ and $M_2 = M$ (bottom left) and the the difference in the best-fit χ^2 per degree of freedom between two models with parameters $M_1 = M+1$ and $M_2 = M$ (bottom right). While the absolute value of the χ^2 can be decreased almost arbitrarily by increasing the parameters M, the improvement at some point ceases to become relevant compared to a model with with $M-1$ parameters, ie the fit does not become better with respect to a model with just one less parameter. Also, the best-fit χ^2 per degree of freedom $N-M$ seems to have

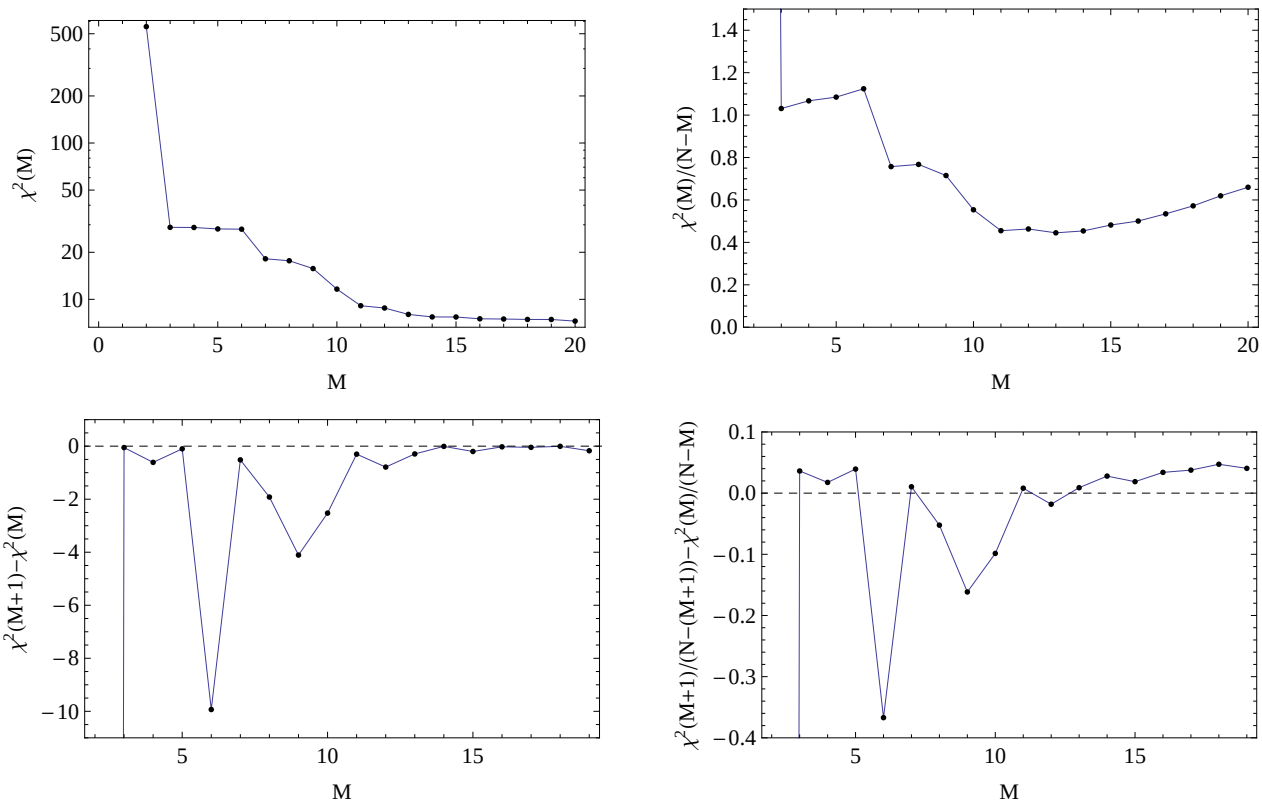

FIG. 1: The best-fit χ^2 as a function of the number of parameters M (top left), the best-fit χ^2 per degree of freedom $N-M$ as a function of the number of parameters M (top right), the difference in the best-fit χ^2 between two models wtih parameters $M_1 = M+1$ and $M_2 = M$ (bottom left) and the the difference in the best-fit χ^2 per degree of freedom between two models with parameters $M_1 = M+1$ and $M_2 = M$ (bottom right).

a minimum, in this case for $M = 13$, which means that beyond that point adding more parameters just does not increase the quality of the fit.

In Fig. 2 we show contour plots of the Bayes factor $\log B_{12}$ of Eq. (3.26), calculated in the case of flat priors by using Eq. (3.18) when the priors are taken for consistency to be proportional to the errors of the best parameters $\Delta a_i = n\ \sigma_i$ for $n = 3$ (left) and $n = 7$ (right). The red color corresponds to a high value for the Bayes factor $\log B_{12}$, ie model M_1 preferred, blue to a low (negative) value for the Bayes factor $\log B_{12}$, ie model M_2 preferred, while white corresponds to equal evidence for both models. The relevant values of the Jeffreys' scale are given in Table II.

As it can be seen, cases that one would normally expect for M_2 to be ruled out, eg $M_1 = 4$ and $M_2 = 14$, as the difference in parameters is a staggering $M_2 - M_1 = 10$ thus giving M_2 a big disadvantage, has a Bayes factor $\log B_{12} = 1.2$ for $n = 3$ and is actually allowed by the Jeffrey's scale! Another similar example can be seen for $M_1 = 4$, $M_2 = 10$ and $n = 7$, see Fig. 2 on the right, where the Bayes factor is $\log B_{12} = 0$ meaning that these two models are totally equivalent! This simple example clearly demonstrates that the Jeffrey's scale is an inadequate tool for model comparison, since it completely fails even in this simple example. Also, the Bayes factor depends heavily on the size of the priors used, since as it can be seen in the two plots of Fig. 2 the results and the conclusions for the two models M_1 and M_2 are very sensitive in the choice of the priors Δa_i, something which is not yet again taken into account by the Jeffreys' scale.

In Fig. 3 we show contour plots of the Bayes factor $\log B_{12}$ of Eq. (3.26), calculated in the case of gaussian priors by using Eq. (3.15) when the priors are taken for simplicity to be proportional to the errors of the best parameters $H_{ii} = n\ \sigma_i^2$ and $H_{ij} = 0$ for $i \neq j$, when $n = 3$ (left) and $n = 7$ (right). The red color corresponds to a high value for the Bayes factor $\log B_{12}$, ie model M_1 preferred, blue to a low (negative) value for the Bayes factor $\log B_{12}$, ie model M_2 preferred, while white corresponds to equal evidence for both models. The relevant values of the Jeffreys' scale are given in Table II. We find similar results as in the case of the flat priors, ie models that should be excluded by the Jeffrey's scale are in fact allowed.

Finally, we have explicitly checked our methodology with other models as well and we get similar results. Specifically,

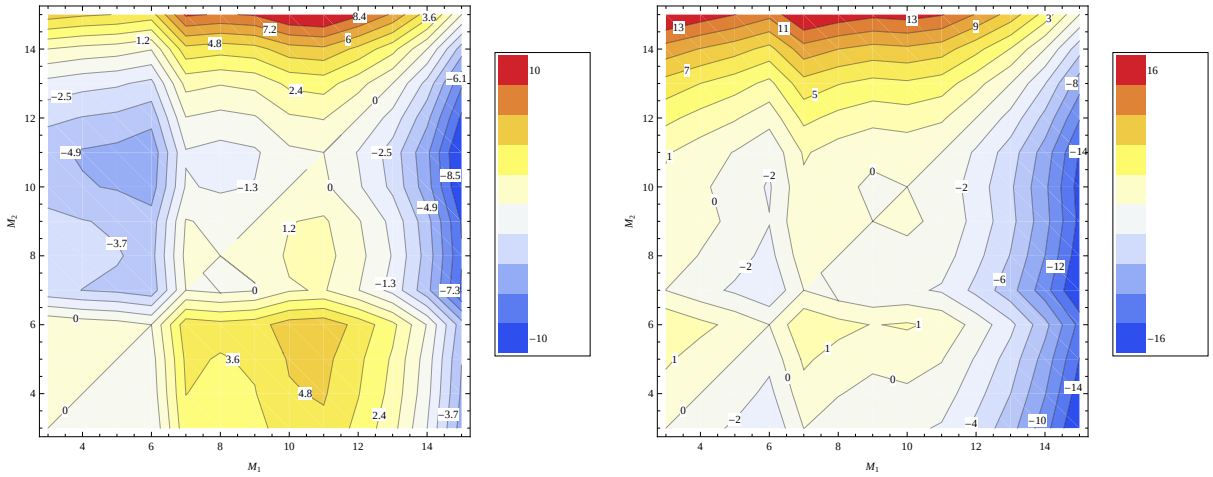

FIG. 2: Contour plots of the Bayes factor $\log B_{12}$ of Eq. (3.26), calculated in the case of flat priors by using Eq. (3.18) when the priors are taken for simplicity to be proportional to the errors of the best parameters $\Delta a_i = n\ \sigma_i$ for $n = 3$ (left) and $n = 7$ (right). The red color corresponds to a high value for the Bayes factor $\log B_{12}$, ie model M_1 preferred, blue to a low (negative) value for the Bayes factor $\log B_{12}$, ie model M_2 preferred, while white corresponds to equal evidence for both models. The relevant values of the Jeffreys' scale are given in Table II.

we also considered the case where the "real" model $f(x)$ is a parabola with three parameters (a, b, c):

$$f(x) = a + b\ x + c\ x^2, \tag{3.29}$$

and we again found the surprising result that cases that one would normally expect for M_2 to be ruled out, eg $M_1 = 4$ and $M_2 = 14$, as the difference in parameters is a staggering $M_2 - M_1 = 10$ thus giving M_2 a big disadvantage, has a Bayes factor $\log B_{12} \sim 1$ for $n = 3$ that is actually allowed by the Jeffrey's scale! So, even when the real model is a function with three parameters, the Jeffrey's scale *fails to rule out* a model with 14 parameters!

To summarize, we firmly believe that the shortcomings of the Jeffreys' scale are the following:

- It is completely insensitive to the size of the priors.
- It fails to work as a "one size fits all" reliable tool for model comparison.
- It provides misleading results even in apparently simple cases, such as the examples considered here.

IV. FIGURE OF MERIT

In this section we will calculate the Figure of Merit (FoM) for this general model by using the usual definition, but we will also introduce a new version which is instead useful for reconstructed quantities that are a function of x.

The $n\sigma$ contours are defined by the constraint equation

$$\mathcal{C} : \chi(\vec{a})^2 = \chi^2_{\rm min} + \delta\chi^2 \tag{4.1}$$

where the value of $\delta\chi^2$ depends on the number of parameters M and the number n of desired σs [16]. This important parameter $\delta\chi^2$ can be found by solving [16]

$$1 - \mathcal{Q}\left(M/2, \delta\chi^2/2\right) = \mathrm{erf}\left(\mathrm{n}/\sqrt{2}\right) \tag{4.2}$$

for $\delta\chi^2 \geq 0$, where $\mathcal{Q}(a, z)$ is the regularized incomplete gamma function $\mathcal{Q}(a, z) \equiv \frac{\Gamma(a,z)}{\Gamma(z)}$ [23]. For various combinations of M and n we show it in Table I. Equation (4.2) can be solved for $\delta\chi^2(M, n)$ as:

$$\delta\chi^2(M, n) = 2\ \mathcal{G}\left(\frac{M}{2}, 1 - \mathrm{erf}\left(\frac{\mathrm{n}}{\sqrt{2}}\right)\right), \tag{4.3}$$

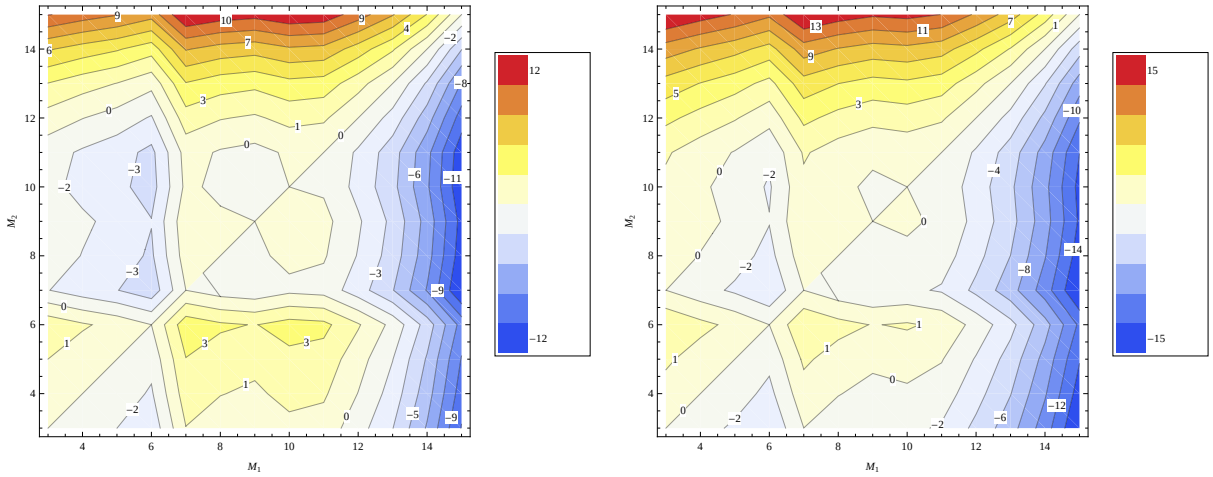

FIG. 3: Contour plots of the Bayes factor $\log B_{12}$ of Eq. (3.26), calculated in the case of gaussian priors by using Eq. (3.15) when the priors are taken for simplicity to be proportional to the errors of the best parameters $H_{ii} = n\ \sigma_i^2$ and $H_{ij} = 0$ for $i \neq j$, when $n = 3$ (left) and $n = 7$ (right). The red color corresponds to a high value for the Bayes factor $\log B_{12}$, ie model M_1 preferred, blue to a low (negative) value for the Bayes factor $\log B_{12}$, ie model M_2 preferred, while white corresponds to equal evidence for both models. The relevant values of the Jeffreys' scale are given in Table II.

where $\mathcal{G}$ is the inverse Γ regularized function[3].

By using Eqs. (2.9) and (2.27)-(2.29) we can rewrite the constraint equation as follows:

$$\begin{aligned} \mathcal{C} : (a - a_{\rm min})_i\ F_{ij}\ (a - a_{\rm min})_j &= \delta\chi^2 \\ \sum_{i=0}^{M-1} s_i^2 &= \delta\chi^2 \end{aligned} \tag{4.4}$$

which in the rotated frame describes a hyper-sphere in M dimensions.

The FoM is equal to the inverse of the volume inside the space whose boundary is given by Eq. (4.1) or in other words the constrained integral

$$\begin{aligned} \text{Vol(M)} &= \int_{\mathcal{C}} d^M a_i \\ \text{FoM} &= \text{Vol(M)}^{-1} \end{aligned} \tag{4.5}$$

This is clearly done so that a smaller volume (better constraints) gives a higher FoM. From Eqs. (4.4) and (2.27)-(2.29) the volume can be expressed as

$$\begin{aligned} \text{Vol(M)} &= \int_{\mathcal{C}} |D|^{-1} d^M s_i \\ &= |F|^{-1/2} V_M(\delta\chi^2) \\ &= |F|^{-1/2} \frac{\pi^{M/2}}{\Gamma(M/2+1)} \left(\delta\chi^2\right)^{M/2} \end{aligned} \tag{4.6}$$

where in the last line we used the fact that the volume of a hyper-sphere of "radius" $R_M = \left(\delta\chi^2\right)^{1/2}$ in M dimensions, which is equal to the constrained integral in the new basis, is $V_M(\delta\chi^2) = \frac{\pi^{M/2}}{\Gamma(M/2+1)} \left(\delta\chi^2\right)^{M/2}$. Finally,

$$\text{FoM}(M) = |F|^{1/2} \frac{\Gamma(M/2+1)}{\pi^{M/2}} \left(\delta\chi^2\right)^{-M/2} \tag{4.7}$$

---

[3] This function can be calculated in Mathematica as $\mathcal{G}(x, y) = InverseGammaRegularized[x, y]$ and works both symbolically and numerically to arbitrary precision.

TABLE III: The determinant of the transformation matrix Λ_{kn} for various combinations of polynomials including the monomials x^n, the Legendre polynomials $P_n(x)$ and the Chebyshev polynomials $T_n(x)$. Clearly, the Chebyshev polynomials provide the best constraints out of all three cases.

$\frac{\mathrm{FoM}_1(M)}{\mathrm{FoM}_2(M)} = \lvert\Lambda\rvert$	Monomials	Legendre	Chebyshev
Monomials	1.000	0.296	0.222
Legendre	3.384	1.000	0.750
Chebyshev	4.511	1.250	1.000

Now, suppose we want to compare two models that have $M_1 = M + \delta M$ and $M_2 = M$ parameters. Then, the ratio of the FoM for the two models can be written as

$$\begin{aligned}\frac{\mathrm{FoM}(M_1)}{\mathrm{FoM}(M_2)} &= \frac{|F^{(1)}|^{1/2}}{|F^{(2)}|^{1/2}} \left(\frac{\frac{\Gamma((M+\delta M)/2+1)}{\pi^{(M+\delta M)/2}} \left(\delta\chi^2\right)^{-(M+\delta M)/2}}{\frac{\Gamma(M/2+1)}{\pi^{M/2}} \left(\delta\chi^2\right)^{-M/2}} \right) \\ &= Z(M,\delta M)\, \frac{|F^{(1)}|^{1/2}}{|F^{(2)}|^{1/2}} \end{aligned} \tag{4.8}$$

where we have defined

$$Z(M,\delta M) \equiv \frac{\frac{\Gamma((M+\delta M)/2+1)}{\pi^{(M+\delta M)/2}} \left(\delta\chi^2(M+\delta M,n)\right)^{-(M+\delta M)/2}}{\frac{\Gamma(M/2+1)}{\pi^{M/2}} \left(\delta\chi^2(M,n)\right)^{-M/2}} \tag{4.9}$$

The dependence of the function $Z(M,\delta M)$ on M and δM for $n = 1$ (1σ) is shown in Fig. 4. As it can easily be seenfrom Eq. (4.9), $Z(M,\delta M)$ is completely independent of the data, depending solely on the number of parameters and the number n of σs. If we want to study the dependence of the FoM of the number of parameters in the case of a nested model, eg when $M_1 = M + 1$ and $M_2 = M$, then it is convenient to define the function

$$\mathrm{Ratio}(M) \equiv \frac{\mathrm{FoM}(M+1)}{\mathrm{FoM}(M)} \tag{4.10}$$

In Fig. 4 we show the dependence of $\mathrm{Ratio}(M)$ on M, but normalized to $\mathrm{Ratio}(M = 2)$. Clearly, adding more parameters does not improve the FoM, when the two models differ just by one parameter, ie $M_1 - M_2 = 1$.

We can also explore the dependence of the FoM on the different bases $X_n(x)$, as shown in Section II B, but now with the same number of parameters M. If we denote the FoM for basis 1 as $\mathrm{FoM}_1(M)$ and the FoM for basis 2 as $\mathrm{FoM}_2(M)$, then by using Eqs. (4.7 and (2.19) we have

$$\frac{\mathrm{FoM}_1(M)}{\mathrm{FoM}_2(M)} = \frac{|F_1|^{1/2}}{|F_2|^{1/2}} = |\Lambda| \tag{4.11}$$

where we have assumed that $F_1 = \Lambda^T\, F_2\, \Lambda$. In the case of our example we have that $x_{max} = 1.55$, so in Table III we show the ratio of the FoM between the different combinations of bases. Clearly, the Chebyshev polynomials provide the best constraints out of all three cases.

V. CONCLUSIONS

We are entering an era where progress in cosmology is driven by data, and alternative models will have to be compared and ruled out according to some consistent criterium. The most conservative and widely used approach is Bayesian model comparison. Naively, one expects the Bayes factor to act as a discriminant among competing models by penalizing those with a larger set of parameters. This has been the common use of Bayesian model comparison in cosmology in the last decade. However, by explicitly computing the Bayes factors for models that are linear with respect to their parameters, we have shown that more information is needed in order to discriminate among models. In particular, we have seen that the thresholds associated to the so called Jeffreys' scale are not as conclusive as most people think they are. We have determined how accurate its predictions are in a simple case where we fully understand and can calculate everything analytically.

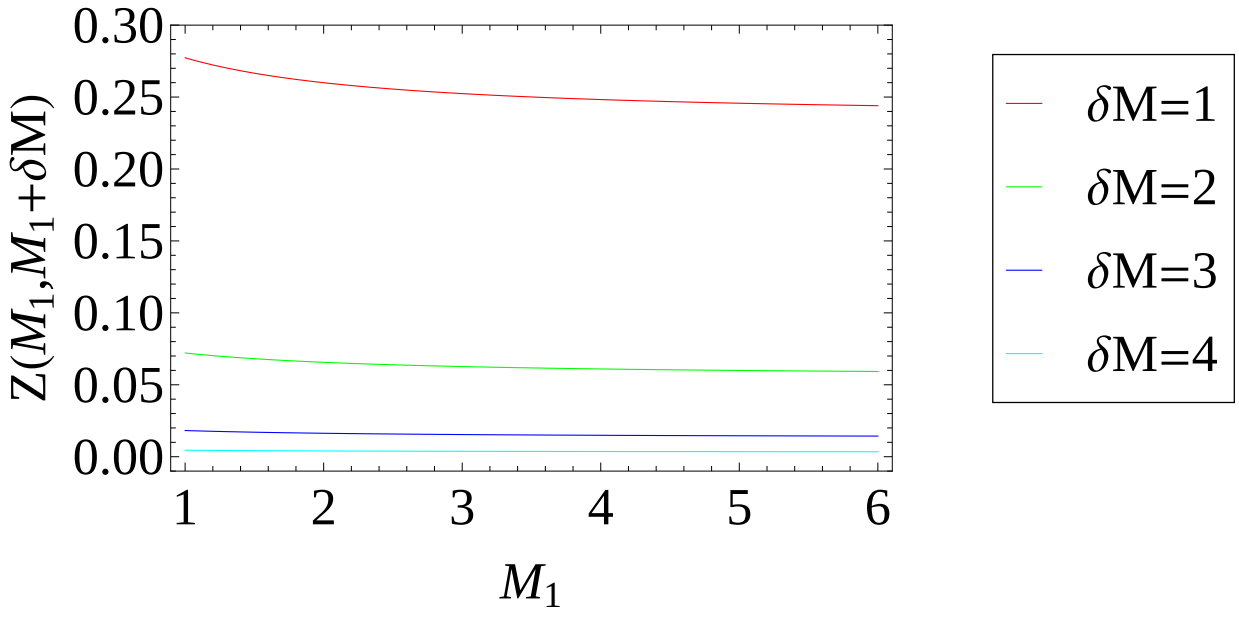

FIG. 4: The dependence of the function $Z(M, \delta M)$ given by Eq. (4.9) on M and δM for $n = 1$ (1σ).

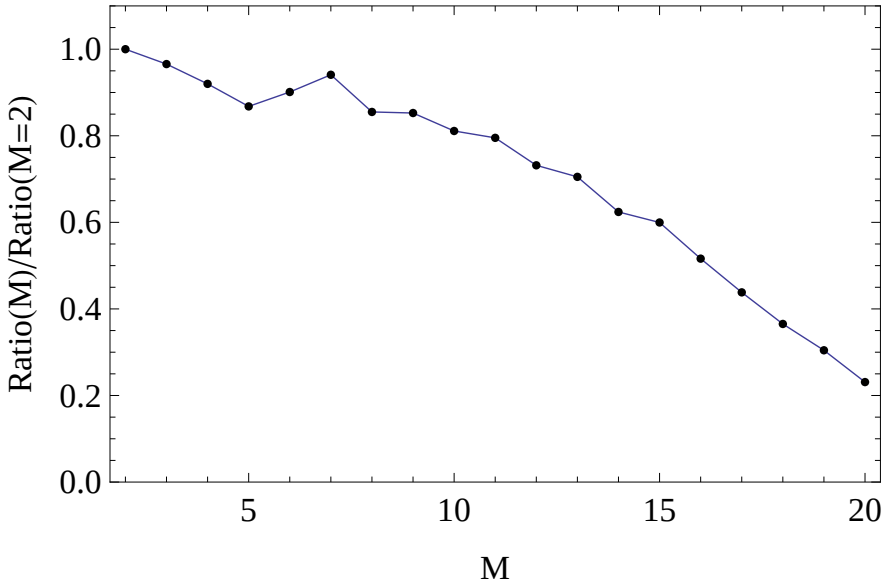

FIG. 5: The dependence of Ratio(M) on M, but normalized to Ratio($M = 2$). Clearly, adding more parameters does not improve the FoM, when the two models differ just by one parameter, ie $M_1 - M_2 = 1$.

We presented our results on a test of the Jeffreys' scale, which is used for comparing models, by explicitly calculating the best fit parameters, the minimum χ^2 and the Bayes factors for all models that are linear with respect to their parameters regardless of the basis used, like in Eq. (2.1). The calculation of the Bayes factor was done for both flat and gaussian priors and analytic formulas were derived in both cases. We also considered the case of changing basis by a transformation from the original *polynomial* basis $X(x)$ with a set of M linearly dependent parameters a_i to a new set with the same number of linearly dependent parameters M and a *polynomial* basis $\tilde{X}(x)$. Furthermore, we also discussed the case of nested models, eg one with M_1 and another with $M_2 \supset M_1$ parameters and we derived analytic expressions for the Bayes factor, while in Section IV we discussed the same problem for the Figure of Merit.

Our main result for the logarithmic Bayes factor between two models that linearly depend on the parameters, Eq. (3.26), does not only contain the difference between the χ^2_{min} of the two models but also contains information

on their covariances. Unsurprisingly, the covariances depend strongly on the data and the model at hand, thus introducing a further complication in model comparison. Therefore, the Bayes factor cannot be the only discriminant among models and thus cannot be taken as a quantitative version of Occam's razor, which simply penalizes models with a larger number of parameters. It is model predictiveness, not model simplicity, which is "rewarded" in Bayesian model comparison [27],[28]. We have shown this by studying analytically the Bayes factors for models with both a small and a large number of parameters to be constrained by the same mock data. In particular, we found that models with $M_1 = 4$ and $M_2 = 14$ parameters had similar Bayes factor ($\ln B_{12}$ 1), and thus were "undecided" by Jeffreys' scale. This could only be understood if the extra 10 parameters do not contribute to the improvement of the model as a description of the data, irrespective of the fact that none of them, M_1 nor M_2, are the "true" model. In fact, even though the one with 14 parameters gets a better χ^2, of course. However, this information is not contained in the Jeffrey's scale, and one could thus be fooled by the Bayes factor to assign similar probabilities to both models.

Another similar example can be seen for $M_1 = 4$, $M_2 = 10$ and $n = 7$, see Fig. 2 on the right, where the Bayes factor is $\log B_{12} = 0$ meaning that these two models are totally equivalent! This simple example clearly demonstrates that the Jeffrey's scale is an inadequate tool for model comparison, since it completely fails even in this simple example. Also, the Bayes factor depends heavily on the size of the priors used, since as it can be seen in the two plots of Fig. 2 the results and the conclusions for the two models M_1 and M_2 are very sensitive in the choice of the priors Δa_i, something which is not yet again taken into account by the Jeffreys' scale.

To conclude, while the Bayes factors are clearly related to the probabilities that one of the two models are more likely than the other, the threshold values of Table II of the Jeffreys' scale used to reject a model in favor of another, are open to interpretation. To make it more clear, the problem is not with the probabilistic interpretation of the Bayes factor, but with the Jeffreys' scale itself. The latter, just represents the threshold after which one is forced to reject a model, usually when $\log B_{12} > 5$ (i.e. strong evidence). What we found was even when one would expect that a model with 14 parameters would and should be ruled out with respect to one with 4 parameters, it was allowed according to the Jeffreys' scale, since $\log B_{12} \sim 1$!

Obviously, having an *ad hoc* scale for model comparison where the thresholds are the same irrespectively of the models and the data, used as a "one size fits all" tool, can lead to *biased* conclusions. The situation can potentially be even worse in cases where the models are not as simple as in the one at hand, e.g. consider the case in cosmology where the models used are non-linear and substantially more complicated [6].

Acknowledgements

We would like to thank A. Liddle and R. Trotta for very useful and enlightening discussions. We acknowledge financial support from the Madrid Regional Government (CAM) under the program HEPHACOS S2009/ESP-1473-02, from MICINN under grant AYA2009-13936-C06-06 and Consolider-Ingenio 2010 PAU (CSD2007-00060), as well as from the European Union Marie Curie Initial Training Network UNILHC PITN-GA-2009-237920. S. N. is supported by CAM through a HEPHACOS Fellowship.

[1] P. A. R. Ade *et al.* [Planck Collaboration], Astron. Astrophys. **536**, 16464 (2011) [arXiv:1101.2022 [astro-ph.IM]].
[2] K. S. Dawson *et al.* [BOSS Collaboration], "The Baryon Oscillation Spectroscopic Survey of SDSS-III," arXiv:1208.0022 [astro-ph.CO].
[3] J. Annis *et al.*, "Constraining Dark Energy with the Dark Energy Survey: Theoretical Challenges," arXiv:astro-ph/0510195. `http://www.darkenergysurvey.org/`
[4] F. R. Bouchet *et al.* [COrE Collaboration], "COrE (Cosmic Origins Explorer) A White Paper," arXiv:1102.2181 [astro-ph.CO].
[5] L. Amendola *et al.* [Euclid Theory WG Collaboration], "Cosmology and fundamental physics with the Euclid satellite," arXiv:1206.1225 [astro-ph.CO].
[6] S. Tsujikawa, arXiv:1004.1493 [astro-ph.CO].
[7] Jeffreys, H., "Theory of probability", Oxford U.P. (1961). Jeffreys, H., "The theory of probability", Oxford U.P. (1998).
[8] Jaynes, E.T., "Probability Theory: the Logic of Science", Cambridge U.P. (2003).
[9] Mackay, D.J.C., "Information theory, inference and learning algorithms", Cambridge U.P. (2003).
[10] D'Agostini, G., "Bayesian reasoning in data analysis: A critical introduction", World Scientific (2003).
[11] Akaike, Hirotugu, "A new look at the statistical model identification", IEEE Transactions on Automatic Control, **16** (1974) 716.
[12] Schwarz, G., "Estimating the dimension of a model", Annals of Statistics **6** (1978) 461.
[13] A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351**, L49 (2004) [astro-ph/0401198].
[14] R. Trotta, Contemp. Phys. **49**, 71 (2008) [arXiv:0803.4089 [astro-ph]].

[15] A. R. Liddle, Ann. Rev. Nucl. Part. Sci. **59**, 95 (2009) [arXiv:0903.4210 [hep-th]].
[16] W. H. Press *et. al.*, "Numerical Recipes", Cambridge University Press (1994).
[17] P. Young, arXiv:1210.3781 [physics.data-an].
[18] M. V. John and J. V. Narlikar, Phys. Rev. D **65**, 043506 (2002) [arXiv:astro-ph/0111122].
[19] Robert, C. P., Chopin, N., & Rousseau, J. 2008, arXiv:0804.3173
[20] R. Lazkoz, S. Nesseris and L. Perivolaropoulos, JCAP **0511**, 010 (2005) [astro-ph/0503230].
[21] C. R. Jenkins and J. A. Peacock, arXiv:1101.4822 [astro-ph.IM].
[22] G. Efstathiou, arXiv:0802.3185 [astro-ph].
[23] Abramowitz, Milton; Stegun, Irene A., eds. (**1972**), "*Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*".
[24] S. Mukherjee, E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley and A. Raftery, [astro-ph/9802085].
[25] S. Nesseris and J. Garcia-Bellido, arXiv:1205.0364 [astro-ph.CO].
[26] C. Gordon and R. Trotta, Mon. Not. Roy. Astron. Soc. **382**, 1859 (2007) [arXiv:0706.3014 [astro-ph]].
[27] M. C. March, G. D. Starkman, R. Trotta and P. M. Vaudrevange, Mon. Not. Roy. Astron. Soc. **410**, 2488 (2011) [arXiv:1005.3655 [astro-ph.CO]].
[28] M. Kunz, R. Trotta and D. Parkinson, Phys. Rev. D **74**, 023503 (2006) [astro-ph/0602378].